arXiv:2507.16389v1 [cs.CV] 22 Jul 2025

From Flat to Round: Redefining Brain Decoding with Surface-Based fMRI and Cortex Structure

Sijin Yu[1] **Zijiao Chen**[2] **Wenxuan Wu**[1] **Shengxian Chen**[1] **Zhongliang Liu**[1]
Jingxin Nie[3] **Xiaofen Xing**[1] **Xiangmin Xu**[1,4] **Xin Zhang**[⋆ 1]
[1]South China University of Technology
[2]Stanford University
[3]South China Normal University
[4]Pazhou Lab

Abstract

Reconstructing visual stimuli from human brain activity (e.g., fMRI) bridges neuroscience and computer vision by decoding neural representations. However, existing methods often overlook critical brain structure-function relationships, flattening spatial information and neglecting individual anatomical variations. To address these issues, we propose (1) a novel sphere tokenizer that explicitly models fMRI signals as spatially coherent 2D spherical data on the cortical surface; (2) integration of structural MRI (sMRI) data, enabling personalized encoding of individual anatomical variations; and (3) a positive-sample mixup strategy for efficiently leveraging multiple fMRI scans associated with the same visual stimulus. Collectively, these innovations enhance reconstruction accuracy, biological interpretability, and generalizability across individuals. Experiments demonstrate superior reconstruction performance compared to SOTA methods, highlighting the effectiveness and interpretability of our biologically informed approach.

1. Introduction

The human brain, shaped by evolution, can be viewed as a highly optimized, naturally "pre-trained" neural network. It encodes sensory stimuli such as visual and auditory signals from various organs into intricate patterns of neural activity. Brain decoding, the reverse of this process, aims to reconstruct sensory stimuli from recorded brain activity, typically measured using functional magnetic resonance imaging (fMRI). This decoding holds significant implications for both neuroscience, by revealing how the brain represents sensory information, and brain-computer interfaces (BCIs), by translating neural signals into actionable outputs.

A pivotal area of brain decoding is fMRI-image reconstruction, which focuses on the exploration of the brain's visual functions. Over time, techniques have evolved from semantic category reconstruction [7, 9, 18] to pixel-level reconstruction [2, 35]. Currently, the most popular approach is to train an fMRI encoder [4, 5, 24, 32, 38, 40, 45], whose output fMRI embeddings are used as conditions to guide a diffusion model [8, 14, 31] for image reconstruction. Following these approaches, fMRI-image reconstruction has expanded from single-subject to cross-subject applications [16, 29, 36, 43, 46]. Methods that encode fMRI signals and use them as conditions to guide diffusion models have achieved success.

However, existing methods share fundamental limitations stemming from the oversimplification of fMRI data as one-dimensional signals, disregarding critical spatial and structural properties. Specifically: (1) fMRI signals arise from voxels located on the cerebral cortex, naturally forming spatial patterns on a two-dimensional non-Euclidean cortical surface [11]. Ignoring this spatial organization diminishes the richness of spatial information available. (2) Structural brain differences among individuals significantly influence functional responses [12, 26], resulting in varied neural responses to identical stimuli [19]. Current methods fail to effectively capture these structural variations, thus impairing performance in cross-subject generalization.

Based on this observation, we propose the Sphere Tokenizer, which processes fMRI signals into fMRI tokens that incorporate spatial structural information. These tokens are then fed into the fMRI encoder for encoding. fMRI signals located on the cerebral cortex are mapped onto a standard sphere using FreeSurfer-based [10] method. Spherical convolution [17, 51] is introduced to extract fMRI features. Similarly to planar convolution [20], each voxel (pixel) only gathers information from its neighbors and updates its value. In addition, unlike previous spherical convolution frameworks, we introduce structural information of the

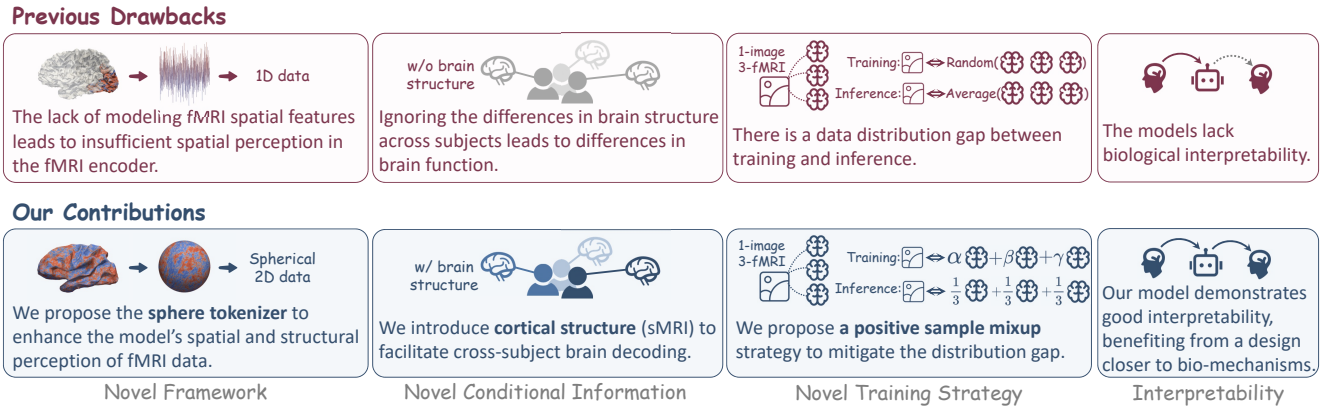


Figure 1. The drawbacks of previous work and the **contributions** of this paper.

cerebral cortex and sphere positional embedding into the Sphere Tokenizer. The cortical structure allows the model to perceive differences in brain structures across subjects, while the sphere positional embedding enables the model to account for functional pattern differences across different regions of the cortex. To the best of our knowledge, we are the first to incorporate brain structure into the fMRI-image reconstruction task.

In addition, current fMRI-image datasets typically provide asymmetric matching samples, *i.e.*, one image corresponds to multiple fMRI scans. Previous work apply each individual fMRI scan for training, but employ the averaged fMRI scan for inference. This creates a data distribution gap between training and inference. To address this, we proposed a positive sample mixup strategy, where during training, we randomly mixup the corresponding fMRI scans to match the averaging process during inference. **Contributions:** we summarize the drawbacks of previous works and the contributions of this paper in Fig. 1.

2. Related Work

fMRI-Image Reconstruction. Early works of vision brain decoding are represented by approaches using sparse linear regression [15], VAE-based methods [9], and GAN-based methods [13, 25, 33]. With the advancement of image generation models, brain decoding tasks have evolved from semantic classification to pixel level reconstruction. With the development of multimodal technologies and diffusion models [8, 14, 31], the current popular pipeline involves aligning fMRI with CLIP [30] and then using it as a condition for diffusion models to perform image reconstruction [4, 32, 38, 40, 45]. Building on this, researchers have extended this pipeline to cross-subject scenarios [16, 29, 36, 43, 46]. These methods typically achieve cross-subject training by introducing subject-specific parameters into the model [29, 43] or incorporating subject-independent tokens [46].

fMRI Tokenizer. The tokenizer processes the raw fMRI data into an appropriate format and passes it to the model. A simple approach is to treat fMRI as 1D data and then process it using a linear module [40] or as one-dimensional convolution [4]. The linear tokenizer and 1D convolution tokenizer simply flatten the fMRI data, without taking its spatial and structural information into account. Therefore, the ROI embedder is proposed [27, 29], which processes fMRI data from different brain regions separately, enabling the fMRI tokens to have better spatial representation capabilities. Unlike the ROI-wise spatial representation of fMRI, our proposed sphere tokenizer provides voxel-wise spatial representation capability.

3. Method

3.1. Data Elaboration and Task Setup

Dataset Our experiments build upon the Natural Scenes Dataset (NSD) [1], which contains large-scale fMRI–image pairs for eight healthy adult participants. The images are sampled from MS-COCO [21]. Following standard practice [29, 32, 37, 43], we focus on four participants (subj01, 02, 05, 07), each having viewed 10,000 images (scanned three times per image). Of these, 1,000 images are shared across all four participants, forming our `test` set. The remaining images are split into 8,500 for `train` and 500 for `val`. We use the fMRI data mapped to the fsaverage surface provided by NSD, then resample it onto the FreeSurfer [10] fsaverage6 standard mesh (40,962 $\times$ 2 voxels per subject). Further preprocessing details are in Appendix §A.

Task Setup and Notations. Our goal is to reconstruct an image I from fMRI responses. Each image I is paired with a COCO annotation T and three corresponding fMRI scans $\{F_i^{\mathrm{L}}, F_i^{\mathrm{R}}\}$ $(i = 1, 2, 3)$, which represent left (L) and right (R) hemisphere activities. The model inputs $F_i^{\mathrm{L}}, F_i^{\mathrm{R}} \in \mathbb{R}^{n\times 1}$ (where n is the number of voxels on a hemisphere),

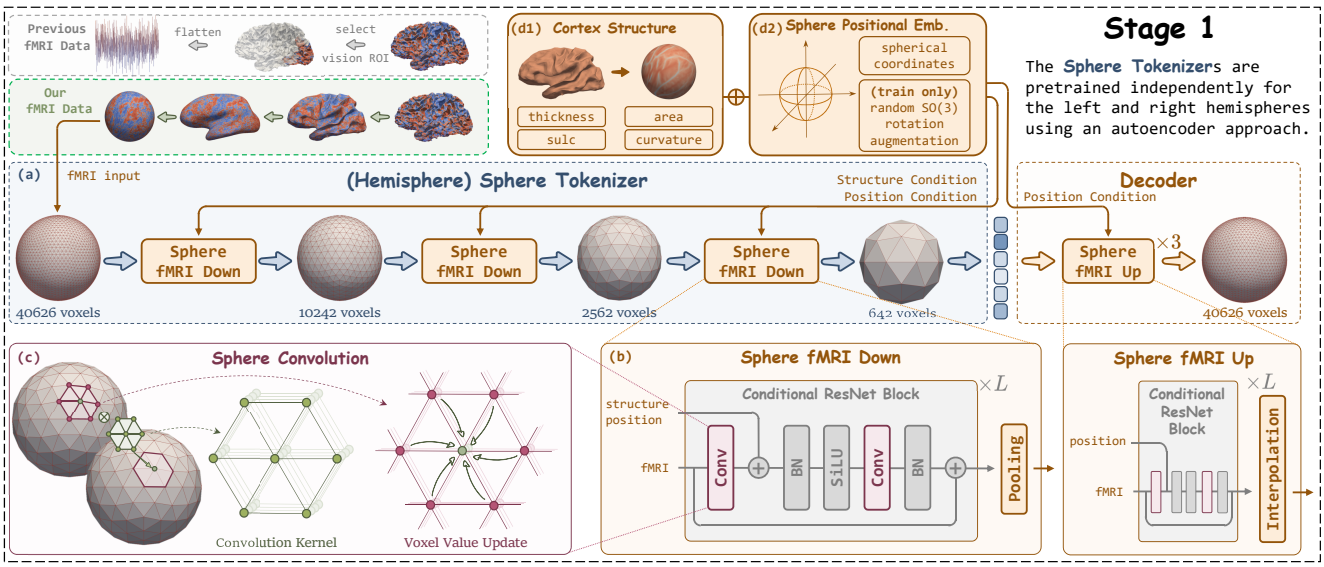


Figure 2. Illustration of sphere tokenizer (§3.2). **(a)** [blue] The tokenizer downsamples the fMRI data three times and maps it to fMRI tokens. **(b)** [brown, bottom] Each downsampling layer consists of L conditional ResNet blocks. **(c)** [purple] The ResNet blocks are based on the sphere convolution, a 2D convolution on the sphere with a receptive field of one-hop neighbors. **(d)** [brown, top] The conditional ResNet block has two types of conditional guidance: cortex structure (d1) and sphere positional embeddings (d2).

and outputs a reconstructed image $\hat{I}$.

To capture both spatial and structural variations in the cortex, we also introduce two additional conditions: cortical structure information $\{S^{\text{L}}, S^{\text{R}}\}$ and sphere positional embeddings P (see §3.2). These elements provide the tokenizer with explicit cues about individual structural differences and topological positions on the cortical surface.

Method Overview. **Stage 1**: A sphere tokenizer (§3.2, Fig. 2) first maps the raw fMRI data to the appropriate fMRI tokens. The sphere tokenizer has strong spatial and structural perception due to (i) data interaction on a 2D spherical surface instead of the previous 1D space, (ii) the incorporation of individual cortical structure information S^{L}, S^{R}, and (iii) spherical position embedding P. **Stage 2**: An fMRI encoder (§3.3, Fig. 3) takes the fMRI tokens output by the sphere tokenizer as input and aligns it with CLIP [30]. **Stage 3**: An image decoder (§3.4, Fig. 3) reconstructs the image from the aligned fMRI embedding.

3.2. Sphere Tokenizer

To leverage the inherent 2D structure of cortical data, we design the sphere tokenizer to project fMRI signals into fMRI tokens. It encodes raw fMRI signals on a spherical surface, then reconstructs them at high resolution to ensure spatial fidelity. We train separate tokenizers for the left and right hemispheres (denoted $\mathcal{T}^{\text{L}}$, $\mathcal{T}^{\text{R}}$), given their roughly mirrored structural layout.

Overview of the Tokenizer. Focusing on the left hemisphere for illustration, the tokenizer $\mathcal{T}^{\text{L}}$ takes high-resolution fMRI data F_i^{L} on fsaverage6 (40,962 voxels), progressively downsamples it three times until fsaverage3 (642 voxels), generating the *fMRI token* x^{L}. A lightweight decoder then upsamples x^{L} back to the fsaverage6 space to compute reconstruction losses (MSE and L1). Formally,

$$x^{\text{L}} = \mathcal{T}^{\text{L}}\big(F^{\text{L}} \mid S^{\text{L}}, P\big), \tag{1}$$

where S^{L} is cortical structure information and P is the sphere positional embedding.

Incorporating Brain Structure and Position. Prior research [12, 26] shows that structural brain properties (*e.g.*, thickness, curvature) strongly influence neural functions. Because no two brains are identical, we introduce structural MRI data to help the tokenizer separate function from individual anatomical variation. In parallel, we add a sphere positional embedding (P) based on 3D coordinates on the cortical surface. This helps the tokenizer distinguish signals arising from different cortical locations.

Sphere Conditional ResNet Blocks. Each downsampling or upsampling stage stacks multiple conditional ResNet blocks, each containing two sphere convolutions [51] that update voxel values using information from their one-hop neighbors on the spherical mesh. Between the two convolutions, we inject structural and positional features to condition the ResNet blocks. Specifically:

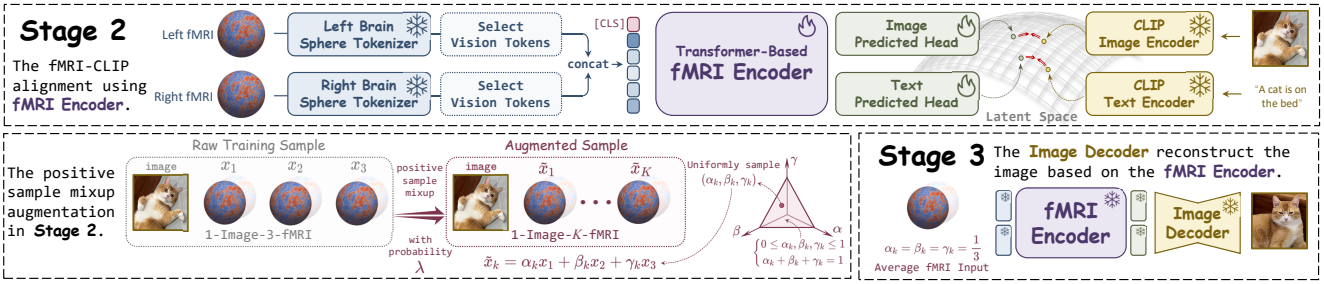


Figure 3. Illustration of the fMRI encoder (§3.3) and the image reconstruction pipeline (§3.4). **(top)** We select vision-relevant tokens from the sphere tokenizer output and feed them into the fMRI encoder, which predicts both image and text embeddings aligned with CLIP. **(bottom left)** Our positive sample mixup strategy: when selected, multiple fMRI scans of the same image are mixed with random weights to better approximate the averaging process used at inference. **(bottom right)** At test time, the three fMRI scans are averaged; the image decoder then reconstructs the final image from the aligned fMRI embedding.

- **Structure Condition** $S^{\mathrm{L}} \in \mathbb{R}^{n\times 4}$: We concatenate four cortical properties (thickness, surface area, sulcal depth, curvature) and resample them to match the current fMRI resolution in the ResNet blocks.
- **Positional Condition** $P \in \mathbb{R}^{n\times 3}$: We project each voxel's 3D coordinates by an MLP, optionally applying random rotations to improve robustness to orientation.

Together, these blocks yield *fMRI tokens* that capture local spatial context (via sphere convolution) and individual cortical distinctions (via structural and positional conditions).

3.3. fMRI-Image-Text Alignment

We apply a transformer-based [42] fMRI encoder ($\mathcal{E}_{\texttt{fMRI}}$), which takes fMRI tokens as input and aligns it with the CLIP [30] (image encoder: $\mathcal{E}_{\texttt{image}}$, text encoder: $\mathcal{E}_{\texttt{text}}$) embedding. The input to $\mathcal{E}_{\texttt{fMRI}}$ is the fMRI vision tokens, which are based on the output of the pretrained sphere tokenizer. The output of $\mathcal{E}_{\texttt{fMRI}}$ is the predicted image embedding $\hat{e}_{\texttt{image}}$ and the predicted text embedding $\hat{e}_{\texttt{text}}$. To more efficiently utilize the 1-image-3-fMRI data in the NSD [1] and alleviate the training-inference data distribution gap, we design a positive sample mixup augmentation.

fMRI Vision Tokens. Once the tokenizer is trained, we can obtain the fMRI token $x^{\mathrm{L}}, x^{\mathrm{R}}$ as a representation of the raw fMRI data. However, in our visual reconstruction task, only the visual brain regions are meaningful. To eliminate redundant fMRI information, we select only the tokens corresponding to the visual brain regions, *i.e.*, $x^{\mathrm{L}}_{\texttt{vis}}, x^{\mathrm{R}}_{\texttt{vis}}$. To isolate visually responsive regions, we apply the "NSDGeneral" ROI mask from NSD [1], selecting only those fMRI tokens corresponding to these visual areas. Then, we concatenate the two hemisphere tokens to obtain the tokens that are input to the fMRI encoder: $x = \texttt{concat}(x^{\mathrm{L}}_{\texttt{vis}}, x^{\mathrm{R}}_{\texttt{vis}})$.

Positive Sample Mixup. In NSD [1], each image is paired with three fMRI scans, *i.e.*, in embedding space, $e_{\texttt{image}}, e_{\texttt{text}} \Leftrightarrow x_1, x_2, x_3$. Here, x_1, x_2, x_3 represent the fMRI vision tokens from the three scans, and "$\Leftrightarrow$" means "corresponds to".

Previous works use $e_{\texttt{image}}, e_{\texttt{text}} \Leftrightarrow x_j$ for training and $e_{\texttt{image}}, e_{\texttt{text}} \Leftrightarrow \texttt{mean}(x_1, x_2, x_3)$ for inference, creating a mismatch between training and testing distributions To resolve this, we propose positive sample mixup, which randomly combines the three fMRI scans for training, reflecting the eventual inference step more accurately. Specifically, we form a mixup sample:

$$\tilde{x}_k = \alpha_k x_1 + \beta_k x_2 + \gamma_k x_3. \tag{2}$$

Here, α_k, β_k and γ_k are random numbers uniformly sampled from the convex polytope that satisfies the linear constraints defined by equation (3).

$$\begin{cases} 0 \le \alpha_k, \beta_k, \gamma_k \le 1 \\ \alpha_k + \beta_k + \gamma_k = 1 \end{cases} \tag{3}$$

Let the number of mixup samples be K. Then, for a given image, we obtain K corresponding fMRI augmentation data samples: $e_{\texttt{image}}, e_{\texttt{text}} \Leftrightarrow \tilde{x}_k$ ($k \in [1:K]$). During contrastive learning, these K samples are used as positive samples with respect to each other (refer to the following "Learning Objectives" part). In inference time, we set $K = 1$, $\alpha_1 = \frac{1}{3}$ and $\beta_1 = \frac{1}{3}$, which is equivalent to $\tilde{x}_1 = \texttt{mean}(x_1, x_2, x_3)$. Our positive sample mixup efficiently leverages the data while mitigating the data distribution gap between training and inference. In practice, we apply positive sample augmentation to a λ proportion of the data within a single batch.

Note. For simplicity, we use the vision tokens x_1, x_2, x_3 to illustrate mixup. In practice, we apply mixup at the tokenizer's input level, *e.g.*, $F^{\mathrm{L}}_1, F^{\mathrm{L}}_2, F^{\mathrm{L}}_3$, to avoid confusion and ensure direct consistency with the rest of our pipeline.

Contrastive Learning. Assume that for a training sample $e_{\texttt{image}}, e_{\texttt{text}} \Leftrightarrow x_k$, the prediction of the fMRI encoder

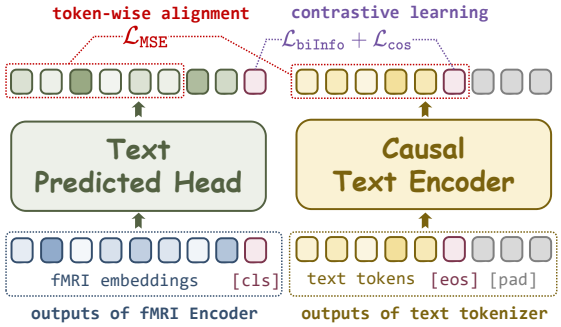


Figure 4. In fMRI-text alignment, we only compute the loss for tokens that carry semantic information, unlike previous work that simply aligns all tokens.

is: $\hat{e}_{\texttt{image},k}, \hat{e}_{\texttt{text},k} = \mathcal{E}_{\texttt{fMRI}}(x_k)$. We compute the bidirectional InfoNCE loss [23] for both the image and the text separately. Taking the image as an example, the bidirectional InfoNCE for sample $\hat{e}_{\texttt{image},k}$ is defined as:

$$\mathcal{L}^{\texttt{image}}_{\texttt{biInfo}} = -\log \frac{K \cdot \exp(\hat{e}_{\texttt{image},k} \cdot e_{\texttt{image}}/\tau)}{\sum_{\texttt{img}=1}^{b} K \cdot \exp(\hat{e}_{\texttt{image},k} \cdot e_{\texttt{img}}/\tau)} - \log \frac{\sum_{k}^{K} \exp(e_{\texttt{image}} \cdot \hat{e}_{\texttt{image},k}/\tau)}{\sum_{\texttt{img}=1}^{b} \sum_{k}^{K} \exp(e_{\texttt{image}} \cdot e_{\texttt{img},k}/\tau)}. \tag{4}$$

Here, b represents the batch size, τ is the temperature hyperparameter, and K is the number of augmentation samples (where `image` is augmented with probability λ) or 3 (when `image` is not augmented with probability $1-\lambda$). The bidirectional InfoNCE for the text $\mathcal{L}^{\texttt{text}}_{\texttt{biInfo}}$ is defined in the same way. The final bidirectional InfoNCE with positive sample augmentation is: $\mathcal{L}_{\texttt{biInfo}} = \mathcal{L}^{\texttt{image}}_{\texttt{biInfo}} + \mathcal{L}^{\texttt{text}}_{\texttt{biInfo}}$. To further reduce the distance between embeddings, we also use a cosine similarity loss:

$$\mathcal{L}^{\texttt{image}}_{\texttt{cos}} = 1 - \frac{\hat{e}_{\texttt{image},k} \cdot e_{\texttt{image},k}}{||\hat{e}_{\texttt{image},k}|| \cdot ||e_{\texttt{image},k}||}. \tag{5}$$

Token-Wise Alignment. Previous work [16, 29, 43] have shown that token-wise alignment is more effective than performing contrastive learning only on the `[cls]` token, *i.e.*, by computing the prediction loss for each token. We adopt this approach for text alignment and further refine it to better suit the textual domain. Because the CLIP [30] text encoder is causal, any tokens generated after the `[eos]` marker in its output embeddings no longer hold meaningful information. As shown in Fig. 4, we only compute the loss for the parts with semantic information in token-wise alignment. For the image, all tokens participate in the loss calculation. The loss for token-wise alignment is MSE ($\mathcal{L}_{\texttt{MSE}}$). Finally, the total loss function is $\mathcal{L} = \mathcal{L}_{\texttt{biInfo}} + \mathcal{L}_{\texttt{cos}} + \mathcal{L}_{\texttt{MSE}}$.

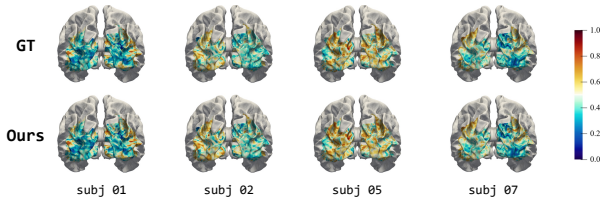


Figure 5. fMRI vision voxels reconstruction results using the sphere tokenizer on the NSD [1] `test`. The sphere tokenizer demonstrates excellent encoding-decoding capabilities across the four subjects. See related analysis in §4.1.

3.4. Image Reconstruction

After obtaining the reconstructed CLIP space semantic embeddings $\hat{e}_{\texttt{image}}$ and $\hat{e}_{\texttt{text}}$, we need an image decoder to reconstruct the image. We choose versatile diffusion [47] as our image decoder, a multimodal diffusion model guided by both CLIP image embeddings and CLIP text embeddings. During inference, we also use classifier-free guidance [31] and Ecphory Mechanism introduced by Psychometry [29].

4. Experiments

Evaluation Metrics. We follow the standard settings in the previous work [29, 32, 37, 43] and use eight image quality evaluation metrics. For low-level properties, we apply pixel-wise correlation (`PixCorr`), structural similarity (`SSIM`) [44], `AlexNet(2)`, and `AlexNet(5)` [20]. For high-level evaluation, we adapt `Inception` [39], `CLIP` [30], `EffNet-B` [41], and `SwAV` [3]. The reported results are the averages computed across four subject.

Reproducibility. The model is trained on one NVIDIA A800 GPU. Inference is conducted on the same machine. Other hyperparameters are at Appendix §C.

4.1. Sphere Tokenizer

The proposed sphere tokenizer maps the raw fMRI data into input embeddings for the encoder, and it is trained using an autoencoder architecture. Since the sphere tokenizer effectively handles data compression before the fMRI encoder, we are interested in its ability to reconstruct fMRI data. We present the fMRI reconstruction results of visual voxels on the NSD `test` using the sphere tokenizer autoencoder (Fig. 5). The displayed data has been min-max normalized to range [0:1] across all visual voxels of each sample. The qualitative results demonstrate a high similarity between the reconstructed fMRI and the groundtruth. This shows that the sphere tokenizer can retain a large amount of informa-

Method	# fMRI Vision Voxels	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Mind-Vis [4]	14K-1D	0.067	0.196	67.7%	74.2%	67.9%	69.3%	0.898	0.513
MindEye [32]	14K-1D	0.129	0.255	84.2%	89.2%	84.1%	85.0%	0.812	0.487
Neuro-Vision [37]	14K-1D	0.265	0.357	93.1%	97.1%	96.8%	97.5%	0.633	0.321
Psychometry [29]	14K-1D	0.297	0.340	96.4%	98.6%	95.8%	96.8%	0.628	0.345
UMBRAE [46]	14K-1D	0.283	0.328	93.9%	96.7%	93.4%	94.1%	0.700	0.393
NeuroPictor [16]	14K-1D	0.141	0.349	91.4%	95.7%	88.3%	88.9%	0.722	0.417
MindBridge [43]	14K-1D	0.151	0.263	87.7%	95.5%	92.4%	94.7%	0.712	0.418
Gu *et al.* [13]	7K-sphere	0.103	0.264	-	-	-	-	0.892	0.508
MindBridge [43]	9K-1D	0.143	0.307	75.0%	86.7%	83.8%	87.0%	0.769	0.413
Ours	9K-sphere	0.165	0.305	78.2%	89.0%	85.1%	88.3%	0.733	0.398

Table 1. Quantitative comparison results on NSD [1] test. All methods are trained using a single model on 4 subjects. Our method uses a lower fMRI resolution (9K) compared to SOTA (14K) due to the standard settings of SphericalUnet [51]. See related analysis in §4.2.



Figure 6. Qualitative results on NSD [1] test subj01. In comparison to previous single-subject models [24, 32, 40, 45] and recent cross-subject models [37, 43], our model is trained on lower fMRI resolution without subject-specific parameters. The visual results of reconstruction demonstrate the effectiveness of our approach. See related discussion in §4.2.

tion from the visual voxels. We also provide more evaluation results for the sphere tokenizer in Appendix §D.

4.2. Comparison to Previous Work

Quantitative Results. We compare our model with eight previous methods: Mind-Vis [4], MindEye [32], Gu *et al.* [13], Psychometry [29], UMBRAE [46], NeuroPictor [16], Neuro-Vision [37] and MindBridge [43]. The quantitative evaluation results are shown in Tab. 1. All models are trained using a single model in the 4 subjects, and the results are the average performance in the 4 subjects. Due to the fact that the sphere tokenizer is based on SphericalUnet [51], which only supports standard spheres, the fMRI resolution we used is different from other studies. We present the fMRI resolution of each model in Tab. 1, expressed in terms of the number of vision voxels. "14K-1D" indicates that each subject has approximately 14K voxels[1] and treats them as a one-dimensional signal. Gu *et al.* applies the spherical convolution framework on the FreeSurfer 32k surface, containing 7,531 visual voxels (7K-sphere). The data we used (9K-sphere) contains 9,548 visual voxels, which entails lower computational costs compared to the 14K data (*e.g.*, in contrast to 7T, our model is also compatible with 3T fMRI data). Compared to methods that also use a spherical framework as the fMRI encoder (Gu *et al.* [13]), our model shows a significant performance improvement (*e.g.*, $0.103/0.264 \rightarrow 0.165/0.305$ in low-

[1] subj01: 15,247, subj02: 14,278, subj05: 13,039, and subj07: 12,682.

#	Setting	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
1	Conv1×1	0.190	0.319	64.0%	78.9%	76.7%	79.4%	0.819	0.405
2	ROI Embedder	0.139	0.312	73.6%	85.4%	83.4%	85.5%	0.771	0.415
3	w/o structure condition	0.146	0.311	75.4%	85.9%	83.3%	86.0%	0.758	0.410
4	w/o positional condition	0.142	0.302	74.0%	85.0%	84.3%	86.0%	0.758	0.413
5	w/o token-wise $\mathcal{L}_{\text{MSE}}$	0.118	0.301	69.6%	79.8%	79.2%	81.6%	0.809	0.448
6	full tokens $\mathcal{L}_{\text{MSE}}^{\text{text}}$	0.152	0.312	75.6%	86.7%	83.9%	86.7%	0.751	0.408
7	Full Model	0.165	0.305	78.2%	89.0%	85.1%	88.3%	0.733	0.398

Table 2. Ablation study (§4.3) on sphere tokenizer on NSD [1] test. The experimental results substantiate the efficacy of the sphere tokenizer, the structural and positional conditions, as well as the token-wise alignment.

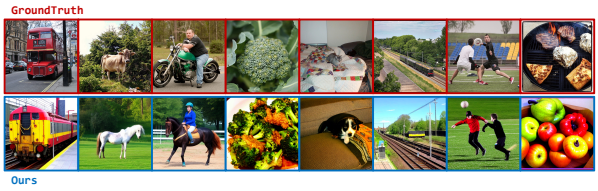


Figure 7. Failed cases in the NSD [1] test subj01. In these cases, although the reconstruction results are not entirely satisfactory, our model has managed to reconstruct partial or analogous semantics. See related analysis in §4.2.



Figure 8. Visual samples on NSD [1] test over 4 subjects. The results demonstrate the cross-subject robustness of our model.

level metrics PixCorr and SSIM, and $0.892/0.508 \rightarrow 0.733/0.398$ in high-level metrics EffNet-B and SwAV). However, compared to the current state-of-the-art, our model performs relatively worse due to the use of a lower fMRI resolution. To demonstrate that this performance decline is solely due to resolution (14K *vs.* 9K), we replicate MindBridge [43] at the same resolution (9K-1D). At the same resolution of fMRI data, our model improves upon MindBridge by $0.022/-0.002/3.2\%/2.3\%$ on the four low-level metrics and by $1.3\%/1.3\%/0.036/0.015$ on the four high-level metrics. Additionally, our model outperforms higher resolution models in certain aspects, *e.g.*, our model shows improvements over MindEye [32] by $1.0\%/3.3\%/0.079/0.089$ on the four high-level metrics.

Qualitative Results. We present some brain decoding samples on the NSD [1] test for subj01 in Fig 6. The baseline models are Takagi *et al.* [40], Brain-Diffuser [24], MindEye [32], DREAM [45], Neuro-Vision [37] and MindBridge [43]. The results of MindBridge are obtained on low fMRI resolution (9K-1D). The image quality reconstructed by our method is not worse than that of other methods. We also provide the failed cases in Fig. 7. In these cases, our model captures partial but incomplete semantics (*e.g.*, riding a motorcycle → riding a horse, cat → dog, and frisbee → soccer), which affects quantitative metrics. We also present the reconstruction results for four subjects in Fig. 8, demonstrating the model's cross-subject robustness. For more visualization results, please refer to Appendix §D.

4.3. Ablation Study

Different Tokenizers. We first evaluate the effectiveness of the proposed sphere tokenizer. We consider two other variants of the fMRI tokenizer: Conv1×1 [4] and ROI Embedder [27, 29]. The results are shown in Tab. 2. The results show that Conv1×1 has poor ability to capture the spatial and structural properties of fMRI data. When replacing the fMRI tokenizer from a Conv1×1 with ROI Embedder that considers the spatial characteristics of different brain regions, we observe a significant performance improvement. When the modeling of fMRI spatial information is upgraded from the ROI level (ROI Embedder, #2) to the voxel level (ours, Sphere Tokenizer, #7), we also observe performance improvement (*e.g.*, the performance increase $4.6\%/3.6\%/2.8\%/2.8\%$ across AlexNet(2), AlexNet(5), Inception and CLIP).

Cortex Structure and Positional Embedding. Tab. 2 also investigates the impact of cortex structure and positional embedding in the sphere tokenizer. As the re-

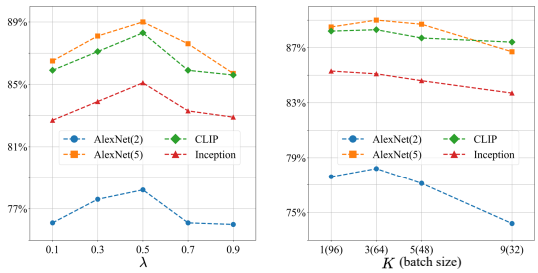


Figure 9. The impact of λ and K in positive sample mixup (§4.3).

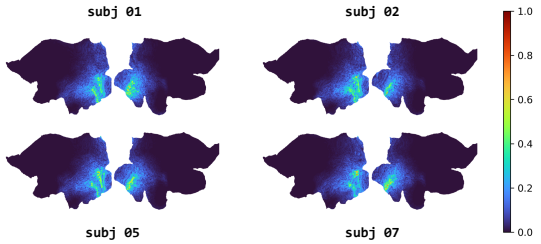


Figure 10. Average neural activity representation differential heatmaps on the NSD [1] `test`. See analysis in §4.4.

sults show, the introduction of the structural information of the cerebral cortex into sphere tokenizer brings a performance boost (*e.g.*, $83.4\%/85.5\%/0.771/0.415 \rightarrow 85.1\%/88.3\%/0.733/0.398$ on all high-level metrics from #3 to #7). This suggests that subject-specific brain structural information facilitates cross-subject brain decoding. Additionally, when removing the sphere positional embedding (#4), we observe a decline in performance.

Token-Wise Alignment. We first remove the token-wise alignment, optimizing only the `[cls]` token, which led to a performance drop (#5). Additionally, we observe that the attention weights for each token are almost identical, causing the transformer to degrade into an MLP. When optimizing all tokens during the fMRI-text alignment (#6), the meaningless `[pad]` tokens in the text also contribute to the loss, leading to a certain degree of performance degradation.

Positive Sample Mixup. We also design ablation experiments to validate the effectiveness of the positive sample mixup. There are two adjustable hyperparameters here: the mixup ratio λ and target mixup samples number K. However, variations in λ and K may affect the batch size. To ensure a fair comparison, we control the expected batch size (or number of images in a batch) to be a constant B in the experiments. Let the batch size in an experiment be b, then the following condition holds: $\lambda Kb+3(1-\lambda)b = 3B$. Then we have $b = \frac{3B}{3+(K-3)\lambda}$. We show the impact of λ and K in Fig. 9. Both λ and K have a peak effect on performance, reflecting the trade-off in positive sample mixup. When λ is too low, the augmentation strength is insufficient. On the other hand, when λ is too high, almost all data become synthetic, affecting the characteristics of the original data. For K, a value that is too small makes the augmentation too weak, while a value that is too large leads to an insufficient number of negative samples within a batch. In practice, our base batch size is $B = 64$. To ensure that b is an integer, we conduct four sets of experiments on K with the optimal $\lambda = 0.5$: $K = 1, 3, 5, 9$ with $b = 96, 64, 48, 32$. Based on the results, we ultimately choose $\lambda = 0.5$ and $K = 3$.

4.4. Biological Interpretability

To further our understanding of the importance of different brain voxels in the vision brain decoding task, we compute the average fMRI heatmaps using Grad-CAM [34] across all samples in the NSD [1] `test` (Fig. 10).

Brain Region. The bright spots in are distributed across the visual regions, with high-level areas being more important than low-level areas. Our model focuses more on high-level areas due to the alignment with CLIP [30] semantics.

Cross Subjects. The model explores subject-independent patterns and features when handling cross subject fMRI data. As shown in Fig. 10, the four subjects have very similar fMRI heatmaps. This explains why our model is able to achieve cross-subject vision brain decoding.

5. Discussion and Conclusion

We approach fMRI-image reconstruction from a novel perspective. We introduce the sphere tokenizer, which treats fMRI signals as 2D data on spherical surface instead of 1D. Considering the individual brain structure-function relationships and differences from cortical locations, we additionally incorporate structure and positional condition for the sphere tokenizer. Furthermore, we use positional embedding to help the model better understand the function of different brain regions. In addition, we design positive sample mixup to more efficiently utilize training data and alleviate the data distribution gap between training and inference. Our model achieves satisfactory performance even under the constraint of fMRI resolution.

6. Acknowledgements

This work is funded by Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004), General Program of the National Natural Science Foundation of China (NSFC-62471185), Guangdong Basic and Applied Basic Research Foundation (2024A1515010180), and Key-Area Research and Development Program of Guangdong Province (2023B0303040001).

References

[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 2, 4, 5, 6, 7, 8, 12, 14, 15, 16, 17, 18

[2] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 5

[4] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023. 1, 2, 6, 7, 14

[5] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *NeurIPS*, 2023. 1

[6] Zhaojin Chen, Sijin Yu, Xuejiao Li, Huirong Lei, Jiyu Qian, Yingxue Yao, Zicong Zheng, Guodong Liang, Xiaofen Xing, Xin Zhang, et al. Sss: Signature-sequence-statistical model for exploring the impact of prenatal depression on newborns' brain using fnirs. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2025. 14

[7] David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2

[9] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023. 1, 2

[10] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. 1, 2

[11] Bruce Fischl, Martin I Sereno, and Anders M Dale. Cortical surface-based analysis: Ii: inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999. 1

[12] Panagiotis Fotiadis, Linden Parkes, Kathryn A Davis, Theodore D Satterthwaite, Russell T Shinohara, and Dani S Bassett. Structure–function coupling in macroscale human brain networks. *Nature Reviews Neuroscience*, 25(10):688–704, 2024. 1, 3

[13] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network. In *Medical Imaging with Deep Learning*, pages 107–118. PMLR, 2023. 2, 6

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[15] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017. 2

[16] Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pages 56–73. Springer, 2025. 1, 2, 5, 6, 14

[17] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical cnns on unstructured grids. *arXiv preprint arXiv:1901.02039*, 2019. 1

[18] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005. 1

[19] Meenakshi Khosla, Gia H Ngo, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22):eabe7547, 2021. 1

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1, 5

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 12

[22] Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024. 14

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5, 13

[24] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. 1, 6, 7

[25] Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. 2

[26] James C Pang, Kevin M Aquino, Marianne Oldehinkel, Peter A Robinson, Ben D Fulcher, Michael Breakspear, and Alex Fornito. Geometric constraints on human brain function. *Nature*, 618(7965):566–574, 2023. 1, 3

[27] Xuelin Qian, Yikai Wang, Yanwei Fu, Xinwei Sun, Xiangyang Xue, and Jianfeng Feng. Joint fmri decoding and encoding with latent embedding alignment. *arXiv preprint arXiv:2303.14730*, 2023. 2, 7

[28] Yixiang Qiu, Hao Fang, Hongyao Yu, Bin Chen, MeiKang Qiu, and Shu-Tao Xia. A closer look at gan priors: Exploiting intermediate features for enhanced model inversion attacks. In *European Conference on Computer Vision*, pages 109–126. Springer, 2024. 14

[29] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 233–243, 2024. 1, 2, 5, 6, 7, 14

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 5, 8

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 5

[32] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 2, 5, 6, 7, 14

[33] Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ Van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018. 2

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[35] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019. 1

[36] Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[37] Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. *Advances in Neural Information Processing Systems*, 37:98083–98110, 2025. 2, 5, 6, 7, 14

[38] Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[40] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023. 1, 2, 6, 7

[41] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[42] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[43] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342, 2024. 1, 2, 5, 6, 7

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[45] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8226–8235, 2024. 1, 2, 6, 7

[46] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pages 242–259. Springer, 2025. 1, 2, 6, 14

[47] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 5

[48] Hongyao Yu, Yixiang Qiu, Hao Fang, Bin Chen, Sijin Yu, Bin Wang, Shu-Tao Xia, and Ke Xu. Calor: Towards comprehensive model inversion defense. *arXiv preprint arXiv:2410.05814*, 2024. 14

[49] Hongyao Yu, Yixiang Qiu, Yiheng Yang, Hao Fang, Tianqu Zhuang, Jiaxin Hong, Bin Chen, Hao Wu, and Shu-Tao Xia. Icas: Detecting training data from autoregressive image generative models. *arXiv preprint arXiv:2507.05068*, 2025. 14

[50] Sijin Yu, Xuejiao Li, Huirong Lei, Yingxue Yao, Zhaojin Chen, Zicong Zheng, Guodong Liang, Xiaofen Xing, Xin Zhang, and Chengfang Xu. Attention-based-features-fusion emotion-guided fnirs classification network for prenatal depression recognition. In *International Workshop on PRedic-*

tive Intelligence In MEdicine, pages 12–23. Springer, 2024. 14

[51] Fenqiang Zhao, Shunren Xia, Zhengwang Wu, Dingna Duan, Li Wang, Weili Lin, John H Gilmore, Dinggang Shen, and Gang Li. Spherical u-net on cortical surfaces: methods and applications. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 855–866. Springer, 2019. 1, 3, 6, 12

[52] Tianqu Zhuang, Hongyao Yu, Yixiang Qiu, Hao Fang, Bin Chen, and Shu-Tao Xia. Stealthy shield defense: A conditional mutual information-based approach against black-box model inversion attacks. In *The Thirteenth International Conference on Learning Representations*, 2025. 14

From Flat to Round: Redefining Brain Decoding with Surface-Based fMRI and Cortex Structure

Supplementary Material

A. Data Preprocessing

fMRI Data. This paper is based on the Natural Scenes Dataset (NSD)[2] [1]. Since we use spherical data as model's input, we employ the data preprocessed with FreeSurfer, which is provided by the official NSD dataset. Taking subj01 as an example, the original fMRI path we use is:

```
nsddata_betas/ppdata/subj01/fsaverage
/betas_fithrf_GLMdenoise_RR/
```

The data values here are single-trial beta weights estimated by applying a general linear model (GLM) to the raw fMRI time series, representing the voxel-wise response and its correlation with visual stimuli. The data provided by the official source are registered to the FreeSurfer standard fsaverage7 surface. We resample the data to a 40,962 sphere to match the compatibility with SphericalUNet [51]. Resampling is performed using the open-source SphericalUNet [51] package. Then, we independently apply zero-score normalization for each voxel within the `train` data of a subject. The `val` and `test` data are zero-centered using the mean and variance from the `train` data. The data split follows the standard setup used in previous works.

Cortex Structure Data. Still taking subj01 as an example, our cortical structural data comes from the path:

```
nsddata/freesurfer/subj01/surf
```

We use four types of structural information: cortical thickness, surface area, sulcal depth, and curvature. Taking the left hemisphere as an example, we use the following four files: `lh.thickness`, `lh.area`, `lh.sulc`, and `lh.curv`. We apply the same method as with the fMRI data to resample it to the fsaverage6 surface. Then, we performed zero-score normalization on each individual file (*i.e.*, each hemisphere of each subject).

Image-Text Pair. All images in the NSD dataset are derived from the MS-COCO [21] dataset. The text annotations we use are from the official MS-COCO[3] dataset.

B. Technical Details

Random Rotation of Sphere Positional Emb. The pseudocode is shown in Algorithm B-1. In the algorithm, we apply Rodrigues' rotation formula, a mathematical tool used for rotating 3D coordinates in 3D Euclidean space.

Algorithm B-1: Random Rotation Augmentation of Sphere Positional Embedding

input : Original spherical coordinates $\mathbf{x} \in \mathbb{R}^{n\times 3}$
Maximum rotation angle $\theta_{\max}$
output: Augmented coordinates $\mathbf{x}' \in \mathbb{R}^{n\times 3}$

get a random rotation axis $\mathbf{v}$
$\phi \sim \mathcal{U}(0, 2\pi)$
$\varphi \sim \mathcal{U}(0, \pi)$
$\mathbf{v} = (v_x, v_y, v_z) = (\sin\phi\cos\varphi, \sin\phi\sin\varphi, \cos\phi)$
get a random rotation angle θ
$\theta \sim \mathcal{U}(0, \theta_{\max})$
apply Rodrigues' rotation formula
$\mathbf{K} = \begin{pmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{pmatrix}, \mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
$\mathbf{R} = \mathbf{I} + \sin\theta \cdot \mathbf{K} + (1 - \cos\theta) \cdot \mathbf{K}^2$
$\mathbf{x}' = \mathbf{x}\mathbf{R}$
return $\mathbf{x}'$

Positive Sample Mixup. We present the pseudocode for positive sample mixup augmentation in Algorithm B-2. To uniformly sample a point $\mathbf{w}$ as the mixup weight in the convex polytope defined by equation (3), we sample a point from the 3-dimensional Dirichlet distribution $\mathcal{D}$ with parameters $\mathbf{a} = (1, 1, 1)$. The N-dimensional Dirichlet distribution is the N-dimensional generalization of the Beta distribution, and its probability density function is given by:

$$p_{\mathcal{D}}(\mathbf{w}|\mathbf{a}) = \frac{1}{B(\mathbf{a})} \prod_{n=1}^{N} w_n^{a_n - 1}. \tag{B-1}$$

Here, $B(n)$ is the normalization factor, ensuring that the probability density function integrates to 1 over the domain:

$$B(\mathbf{a}) = \frac{\prod_{n=1}^{N} \Gamma(a_n)}{\Gamma\left(\sum_{n=1}^{N} a_n\right)}. \tag{B-2}$$

[2] https://naturalscenesdataset.org
[3] http://images.cocodataset.org

Algorithm B-2: Positive Sample Mixup

input : Three fMRI scans $\mathbf{x} = (x_1, x_2, x_3)$
Mixup ratio λ
Mixup number K
output: Augmented samples $\tilde{x}_1, \tilde{x}_2, \cdots$

$t \sim \mathcal{U}(0, 1)$
if $t < \lambda$ **then**
  for $k \leftarrow 1$ **to** K **do**
    $\mathbf{a} = (1, 1, 1)$
    $\mathbf{w} \sim \mathcal{D}(\mathbf{a})$ # Dirichlet Dist., Eq. (B-1)
    $\tilde{x}_k = \mathbf{w} \cdot \mathbf{x}$
  return $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_K$
else
  return x_1, x_2, x_3

Multi Positive Samples InfoNCE. We modify the InfoNCE [23] to accommodate the scenario where multiple positive samples exist within a batch. The algorithm for multi positive samples InfoNCE is in Algorithm B-3.

Algorithm B-3: $\mathcal{L}_{\texttt{Info}}$ for Multi Positive Samples

input : Query $\mathbf{q} \in \mathbb{R}^n$
Positive key $\mathbf{k} \in \mathbb{R}^n$
Temperature coefficient τ
Image index $\mathbf{I} \in \mathbb{N}^n$
output: The InfoNCE loss l

$\mathbf{q} = \mathbf{q}/||\mathbf{q}||$
$\mathbf{k} = \mathbf{k}/||\mathbf{k}||$
$\mathbf{W} = \mathbf{q} \cdot \mathbf{k}^\top / \tau$ # $\mathbf{W} \in \mathbb{R}^{n \times n}$
Let sample mask $\theta_{ij} = \begin{cases} 1 & \text{if } \mathbf{I}_i = \mathbf{I}_j \\ 0 & \text{if } \mathbf{I}_i \neq \mathbf{I}_j \end{cases}$
$\mathbf{e}_i^{\texttt{all}} = \log \sum_{j=1}^{n} \exp(\mathbf{W}_{ij})$ # $\mathbf{e}^{\texttt{all}} \in \mathbb{R}^n$
$\mathbf{e}_i^{\texttt{pos}} = \log \sum_{j=1}^{n} \theta_{ij} \exp(\mathbf{W}_{ij})$ # $\mathbf{e}^{\texttt{pos}} \in \mathbb{R}^n$
$l = \sum_{i=1}^{n} (\mathbf{e}_i^{\texttt{all}} - \mathbf{e}_i^{\texttt{pos}}) / n$
return l

C. Training Details

Sphere Tokenizer. The training is conducted on a single NVIDIA A800 80GB GPU. The hyperparameters are shown in Tab. C-1. We use a cosine learning rate scheduler during training. To improve the robustness of the model, we introduce two data augmentation techniques: mixup with parameters `ratio=0.3` and `beta=0.3`, and the random rotation augmentation (maximum rotation angle: $5°$) for position condition mentioned in our paper §3.2. To make the model focus more on visual brain regions, we only compute the loss for visual voxels.

Hyperparameters	Value
encoder channels	[64, 128, 256]
dncoder channels	[32, 64, 128]
hidden channels	32
encoder ResNet blocks per down layer	4
decoder ResNet blocks per down layer	2
batch size	32
learning rate	4.0e-5
weight decay	0.05
max gradient norm	0.1
epoch	80

Table C-1. Hyperparameters for training sphere tokenizer.

fMRI Encoder. The training is conducted on a single NVIDIA A800 80GB GPU. The hyperparameters are shown in Tab. C-1. We use a cosine learning rate scheduler during training. To improve the model's generalization ability, we applied augmentation to the images. Each training image has a 50% chance of being randomly horizontally flipped. The color properties are perturbed as follows: brightness, contrast, and saturation with a maximum perturbation of 0.2, and hue with a maximum perturbation of 0.1. Then the image is randomly rotated by up to $30°$ and scaled within the range of [0.8, 1.0]. The temperature coefficient of $\mathcal{L}_{\texttt{biInfo}}$ is 0.1. To avoid the influence of outliers, the CLIP embeddings are clamped to the range [-1.5, 1.5].

Hyperparameters	Value
embedding dim	1024
MLP ratio	4
depth	24
num heads	16
projection dropout	0.5
batch size	64
learning rate	5.0e-4
weight decay	0.05
max gradient norm	0.5
epoch	30

Table C-2. Hyperparameters for training fMRI encoder.

Inference	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
subj01	0.172	0.314	78.6%	88.7%	84.8%	88.9%	0.736	0.396
subj02	0.167	0.302	77.7%	89.0%	85.9%	88.2%	0.733	0.394
subj05	0.163	0.305	78.6%	90.1%	86.4%	89.6%	0.717	0.393
subj07	0.157	0.298	78.0%	88.3%	83.2%	86.7%	0.746	0.409
Average	0.165	0.305	78.2%	89.0%	85.1%	88.3%	0.733	0.398

Table D-1. Quantitative results for each subject on NSD [1] `test`.

D. More Results

Sphere Tokenizer. We present more reconstruction results of the sphere tokenizer autoencoder in Fig. D-1, which demonstrates its effectiveness.

Comparison to Previous Work. We report the quantitative evaluation metrics in Tab. 1. The results for MindVis [4] and MindEye [32] are cited from the report in [29], while the result for UMBRAE [46] is from the report in [37]. The authors of NeuroPictor [16] fine-tune the model for each subject to achieve higher performance. However, to ensure a fair comparison with other methods, we cite and report the version w/o fine-tuning. The remaining methods are cited from their respective original papers.

Quantitative and Qualitative Results. We present the quantitative results on each subject of NSD [1] `test` in Tab. D-1. We also provide additional decoding results in Fig. D-2, D-3, and D-4. Readers can download all reconstruction images results from here: https://huggingface.co/datasets/yusijin/sphere_tokenizer_results_NSD.

E. Limitations and Future Work

Although we propose a novel approach for vision brain decoding, there are limitations that should be acknowledged.

Low fMRI Resolution. Our model is trained on low fMRI resolution, constrained by the spherical resolution supported by the standard SphericalUNet. Experimental results demonstrate that fMRI resolution has a significant impact on performance. In the future, we plan to (1) use higher-resolution spherical convolutions and (2) adjust the architecture of the sphere tokenizer to more efficiently handle low-resolution fMRI data.

Constrained Dataset Size. Similar to most previous work, we have only validated our results on the NSD dataset. Although current techniques can handle cross-subject decoding, they are far from achieving cross-dataset generalization. To achieve more robust vision brain decoding, larger-scale datasets are required.

Challenges in Practical Application. Our model is based on fMRI, which requires stringent conditions and high costs for data collection. This limits the widespread adoption and application of brain decoding technology. Some technologies that are more suitable for real-time brain activity recording, such as EEG [22] and fNIRS [6, 50], fall far short of fMRI decoding performance due to their low signal-to-noise ratio. This highlights the need for more efficient methods to enable models to capture representations of brain activity.

Challenges in Neuroscience. The scientific community still has no definitive understanding of the detailed mechanisms behind the functioning of the human brain. The development of brain decoding can provide novel perspectives on this, highlighting the significance of biological interpretability in brain decoding models. We will conduct a deeper exploration of this.

Privacy and Security Considerations. Personal brain activity data is highly sensitive private information, so protecting data security is crucial. For example, membership inference attacks [49] and model inversion attacks [28, 48, 52] can cause serious privacy leaks during the inference stage of the model. We will incorporate considerations of data security in the future.

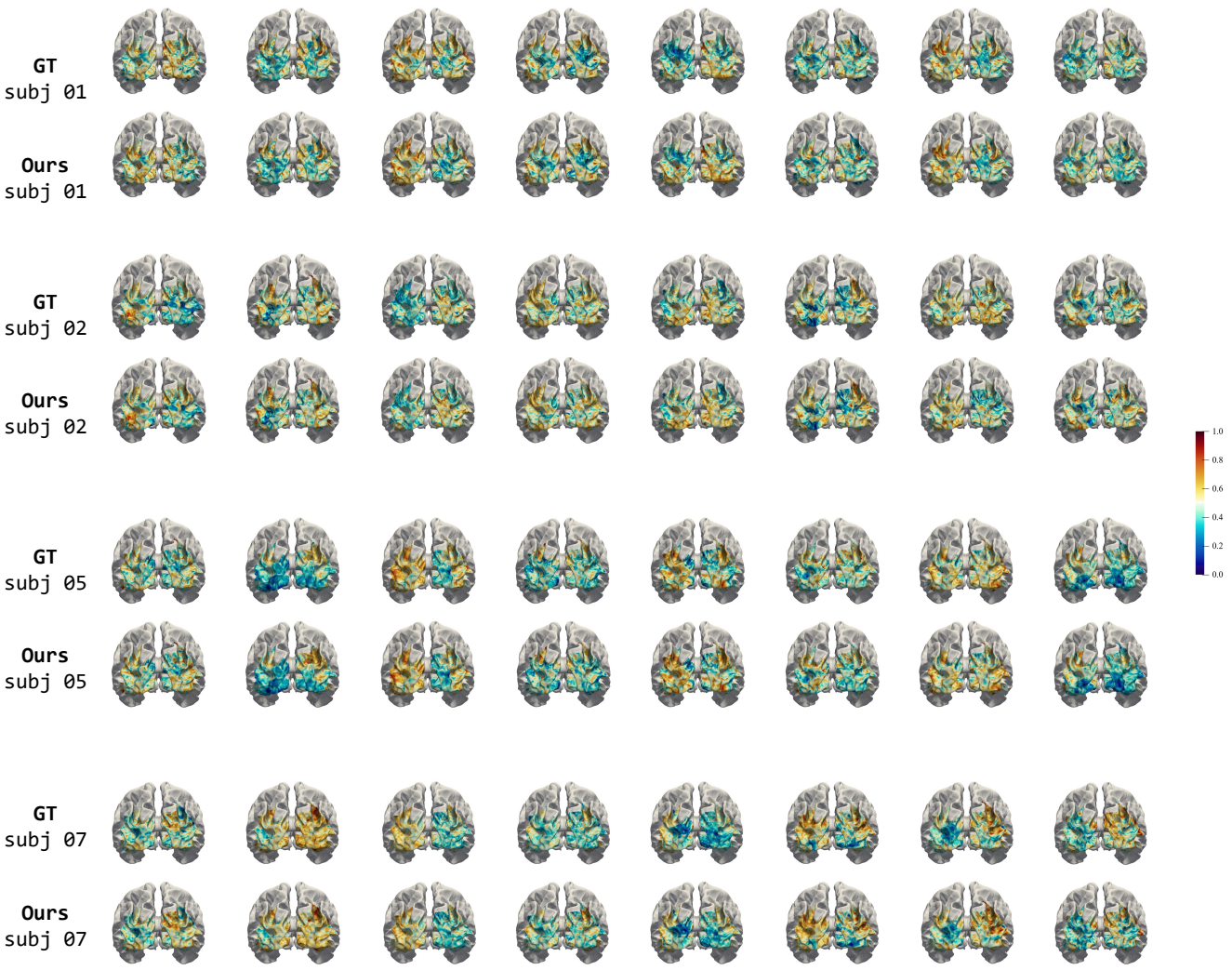


Figure D-1. More fMRI vision voxels reconstruction results using the sphere tokenizer on the NSD [1] `test`.



Figure D-2. More fMRI-image reconstruction results on the NSD [1] `test`.

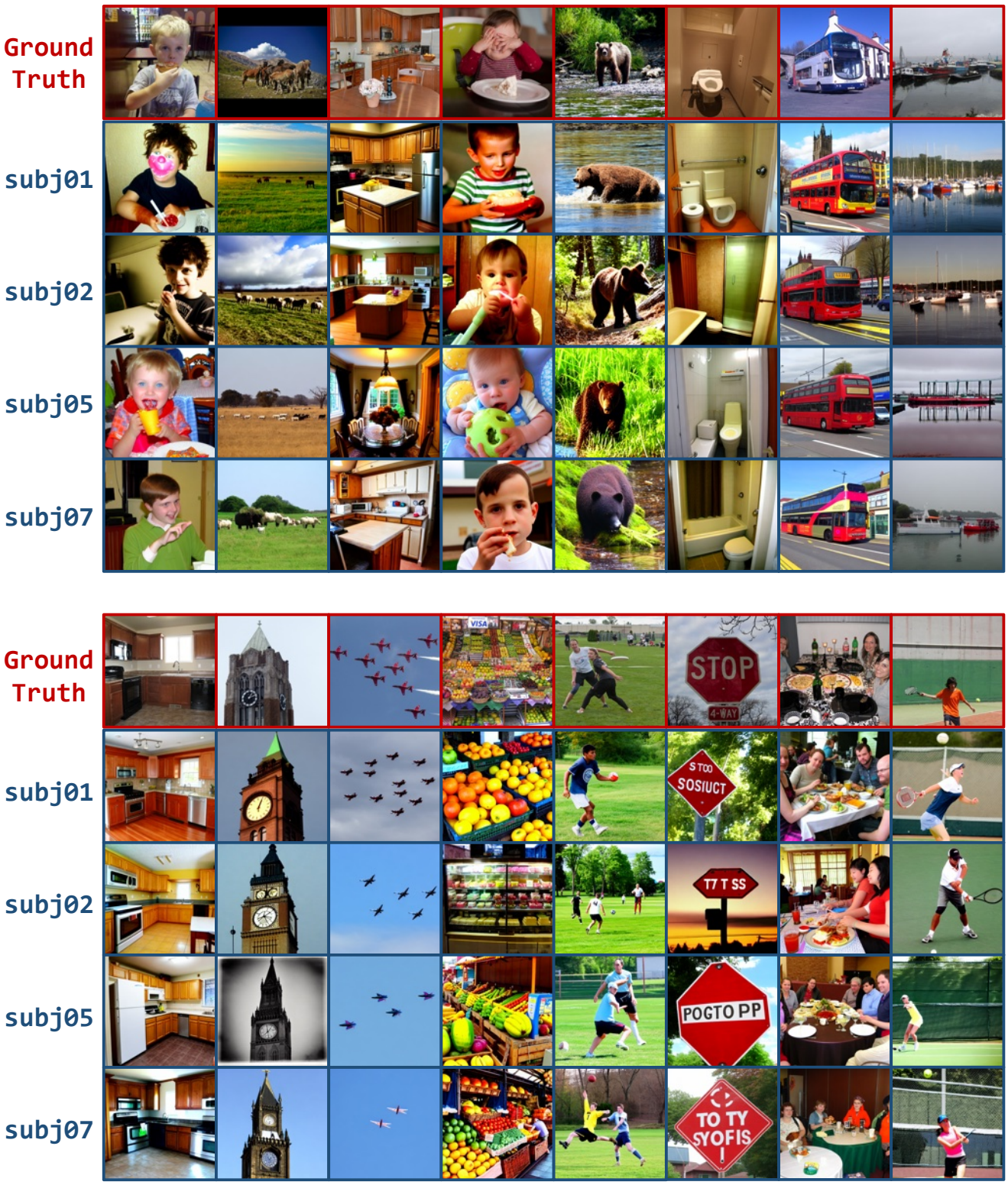


Figure D-3. More fMRI-image reconstruction results on the NSD [1] `test`.



Figure D-4. More fMRI-image reconstruction results on the NSD [1] `test`.